

教師なし学習を用いたCOVID-19新規感染者数のクラスター分析：続報

Cluster Analysis on the Number of COVID-19 Newly Infected People using Unsupervised Learning : a follow-up study

小松 隆行*

Takayuki Komatsu

概要

本稿では、インターネット上に公開されている COVID-19（新型コロナウイルス）に関するオープンデータに関して、相関分析とクラスター分析を行う。なお、前回の報告以降に発生した感染に関する新データと、それを加えた約1年間分の全データについて分析を行う続報となっている。対象とするデータは、日本の全都道府県、日本を含むアジア諸国、およびヨーロッパ諸国の新規感染者数の日次データである。これらを各々時系列データと捉え、地域間の時系列データを統計学的に相関分析し、さらに機械学習の教師なし学習を用いてデータサイエンス的な視点から似たものを同一グループに分類するクラスター分析を試みる。

1. はじめに

COVID-19（新型コロナウイルス）のパンデミックは、2020年1月頃から約3年を経た現在も継続しているが、ワクチン接種など多くの施策が講じられているにもかかわらず、感染者の急拡大と収束を繰り返している。世界各国の感染状況は、あらゆるメディアで報じられており、日次の新規感染者数をはじめとする様々な定量的データも、国毎や自治体毎に、多くのインターネット上のサイトで公開され、ほぼ毎日更新されている。世界各国に関するデータは、オープンデータとして世界保健機関（WHO）のサイト⁽²⁾などで公開されている。日本でも厚生労働省のサイトや各自治体のサイトにおいて、最新のデータが公開されている。日本において2022年の1年間では、感染者数が急増する「波」が3回発生し、都道府県毎のそれらの波の新規感染者数の規模は、2021年までの数倍に拡大し、約10倍に及んでいる場合もある。このことは、前述のサイトで可視化されたグラフなどでも確認できるが、それらのオープンデータをCSV形式やExcel形式でダウンロードし分析や解析を行うことも可能である。その種類によっては欠損などがあるが、基本的な統計分析の結果や知見を、世界の国毎、地域毎、北海道内の大都市や振興局単位で解析することを可能にしている。

本報告の目的は、2022年に発生したデータを追

加して、前回の報告と同様に COVID-19 の新規感染者数のデータをある非定常の情報源から得られる実測値データと考え、これらの時系列データのみから統計学的手法とデータサイエンスの機械学習の手法を用いることによって何らかの知見を得ることである。具体的には、相関分析や機械学習の教師なし学習と呼ばれている手法の Time Series k-means 法⁽³⁾によるクラスタリング（自動グループ分け）を試みて考察を行うことである。

2. 全国47都道府県の新規感染者数に関する分析

2.1 全国47都道府県間の相関分析

まず、日本全国の47都道府県毎のデータ⁽¹⁾について相関分析を行う。前回の報告^(xx)では第1波から第4波の期間、2020/1/16～2021/6/28のデータを用いて、都道府県間の相関係数を算出し、その値のヒートマップとして示した。今回は、感染者数の規模が数倍に激増した、第5波から第8波の相関係数のヒートマップ（デンドログラム）を図1に示す。図1の縦軸と横軸には、47都道府県名が配置され、座標の交差した場所に算出された相関係数が記載されている。これらは、その所在地が概ね北から南の順に地方毎に記載されている。この図1では、相関係数の大小により該当するマス目の色が青色（相関係数 0.5）から赤色（相関係数 1.0）まで変化し

て表示されている。相関係数が 0.9 以上場合は、2 つの変数間に非常に強い相関があると言われていたが、図 1 のヒートマップでは、濃い赤色のマス目の部分が非常に強い相関があると言える。また、相関係数が 0.7 以上 0.9 未満の場合は強い相関があると言われており、図 1 では黄緑から橙色にかけての部分である。デンドログラムで近い位置にグルーピングされ、赤色のマス目が集中しそれらの周囲にも

橙色のマス目が集中している場所としては、北海道と東北地方の 6 県と新潟と長野、東京と大阪、東京の隣接県、大阪の隣接県、東海から近畿、山陽、九州にかけての大都市周辺地域、関東地方の大都市周辺地域が挙げられ、これらの地域間の相関係数の値が特に高いことが分かる。また沖縄県は、他の都道府県との相関が比較的低いことが分かる。

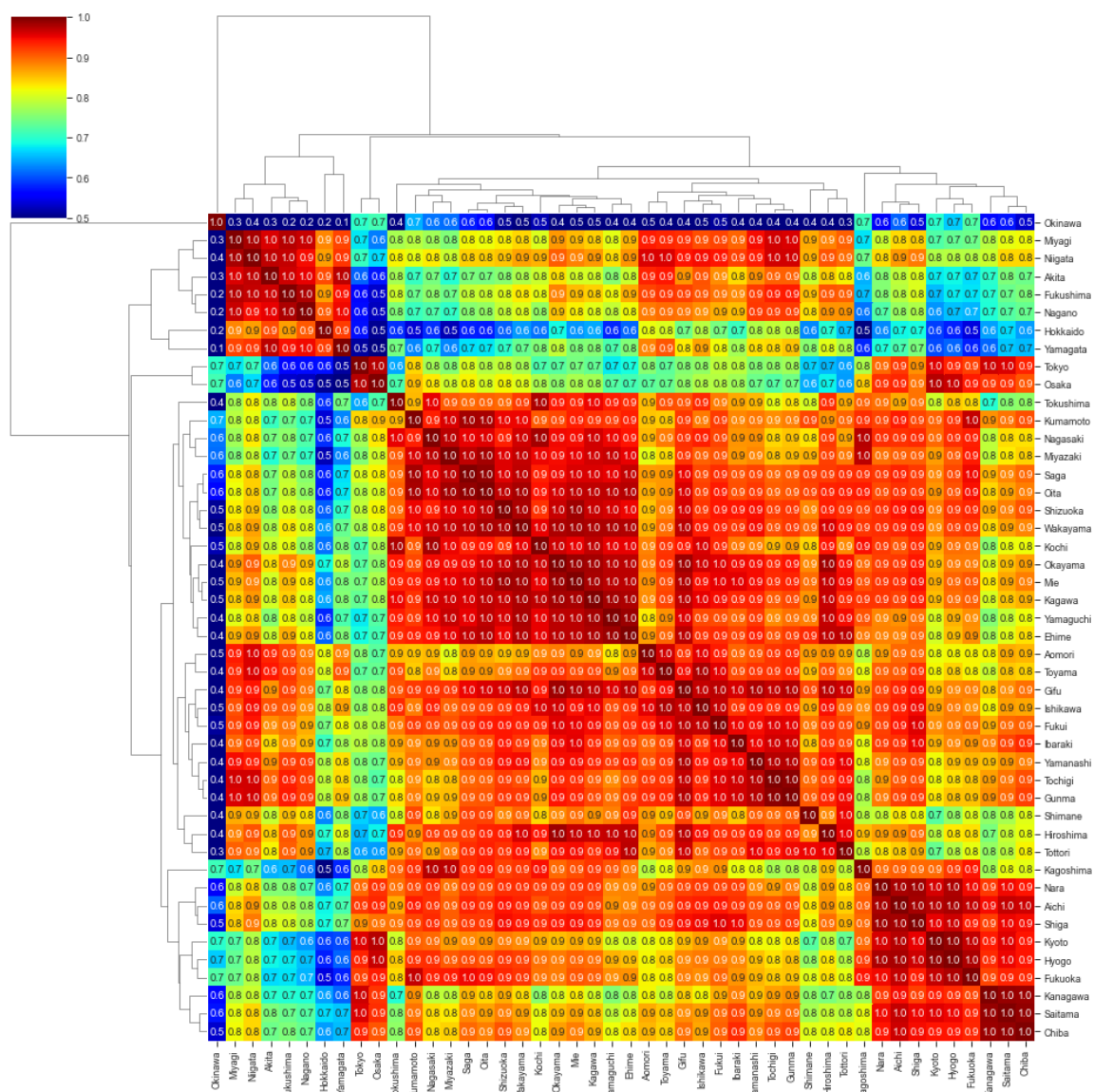


図 1 47 都道府県間の 7 日平均の相関係数ヒートマップ(第 5~8 波:2022 年 1 月 1 日~2023 年 2 月 1 日)

2.2 全国 47 都道府県のクラスタリング

次に、全国 47 都道府県毎に 2020 年 1 月 16 日から 2023 年 1 月 26 日現在までのデータにおいて、平均と標準偏差で標準化された新規感染者数の 14 日平均の推移を Time series k-means 法⁽³⁾ (Euclid 計量)を用いてクラスタリングを行った。クラスタ

数は 13 とした。図 5 は、同じクラスターに分類された都道府県毎の推移を、クラスター毎にプロットしたグラフである。横軸は日付ですべて同一であり、縦軸は標準化による標準偏差の何倍かを表している。それに対応してスケールと最大値がクラスター毎に異なることに留意する。図 2 から

東京と大阪、関東や中部の大都市からやや遠い地域、東海から瀬戸内海沿岸の大都市を除く地域、東北と新潟と長野、愛知と近畿の大阪周辺と福岡、九州の大都市を除く地域、東京の隣接県、山陰、北陸と青森、秋田と山形、北海道、沖縄がそれぞれ単独でクラスタリングされている。大都市を中心にした距離や位置の近さに関するクラスターや、人の往来の量や交通量の大小が関係していると推測されるクラスターが形成されている。

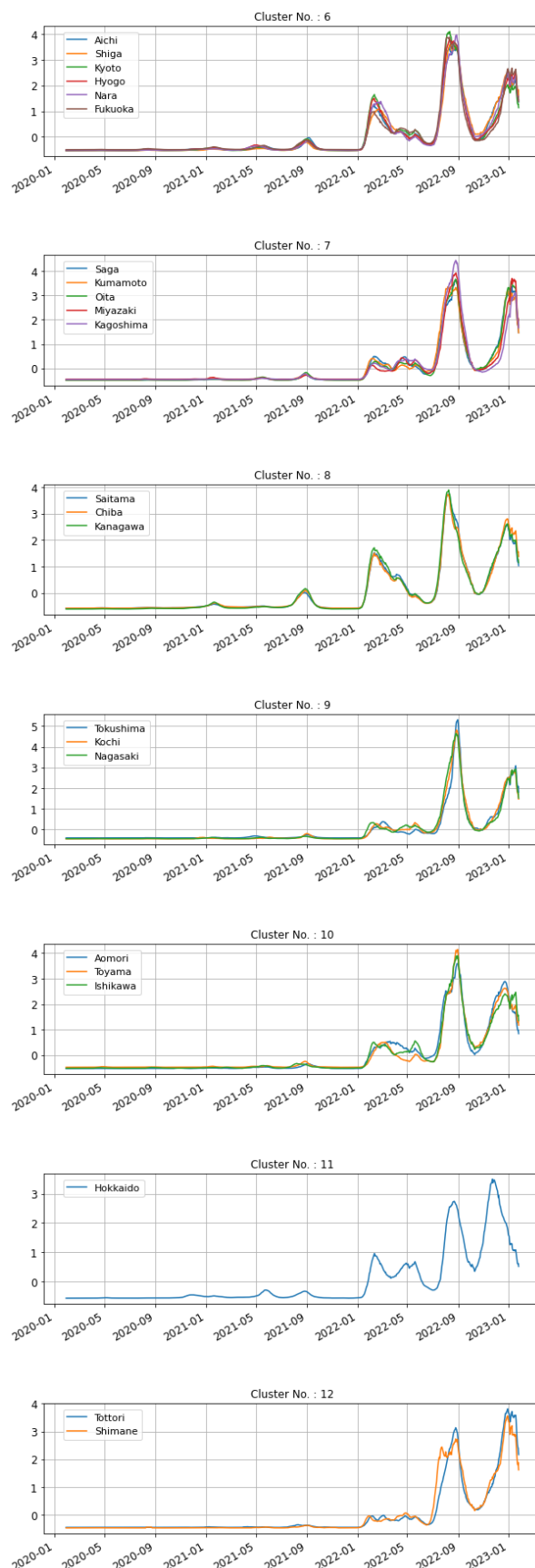
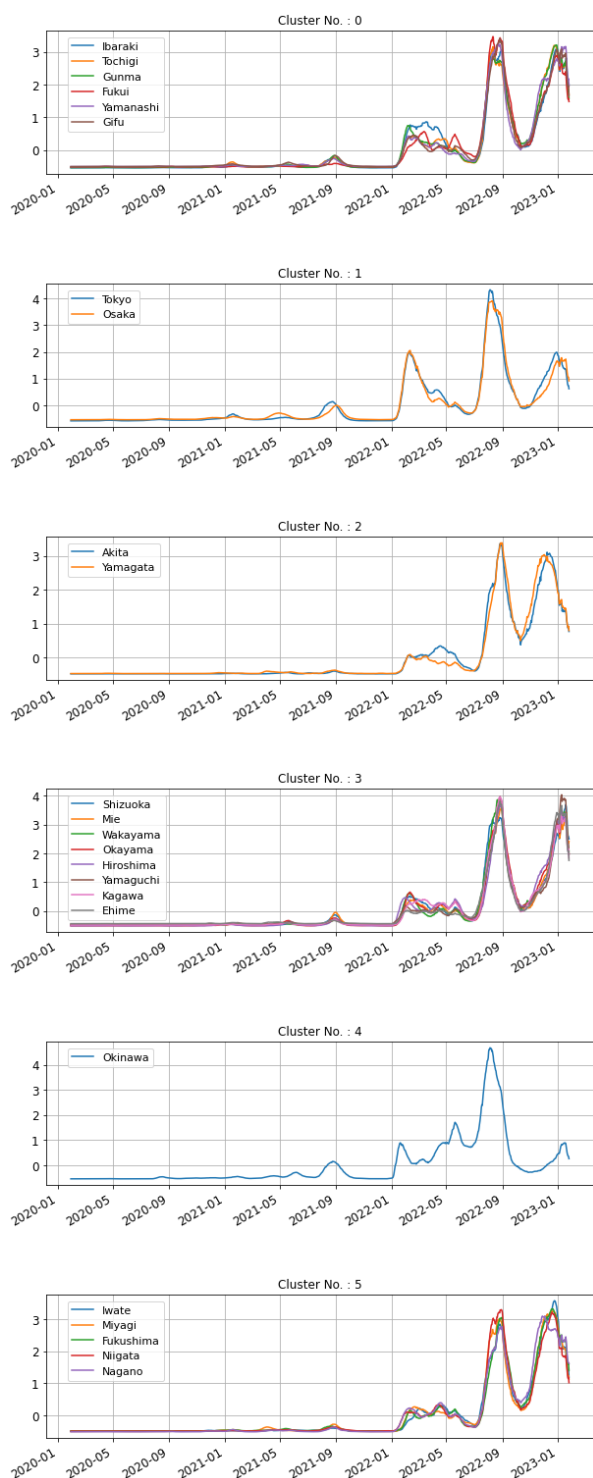


図2 標準化された新規感染者数14日平均のクラスタリング（2020/1/16～2023/1/26）

次に、全国47都道府県毎の10万人当たりの新規感染者数の14日平均の推移を、Time Series k-means 法⁽³⁾ (Euclid 計量) を用いてクラスタリングを行った結果を図3に示す。クラスター数はエルボー法などにより13とした。図2に示した標準化された新規感染者数14日平均のクラスタリングの結果と同様の結果が確認できる。

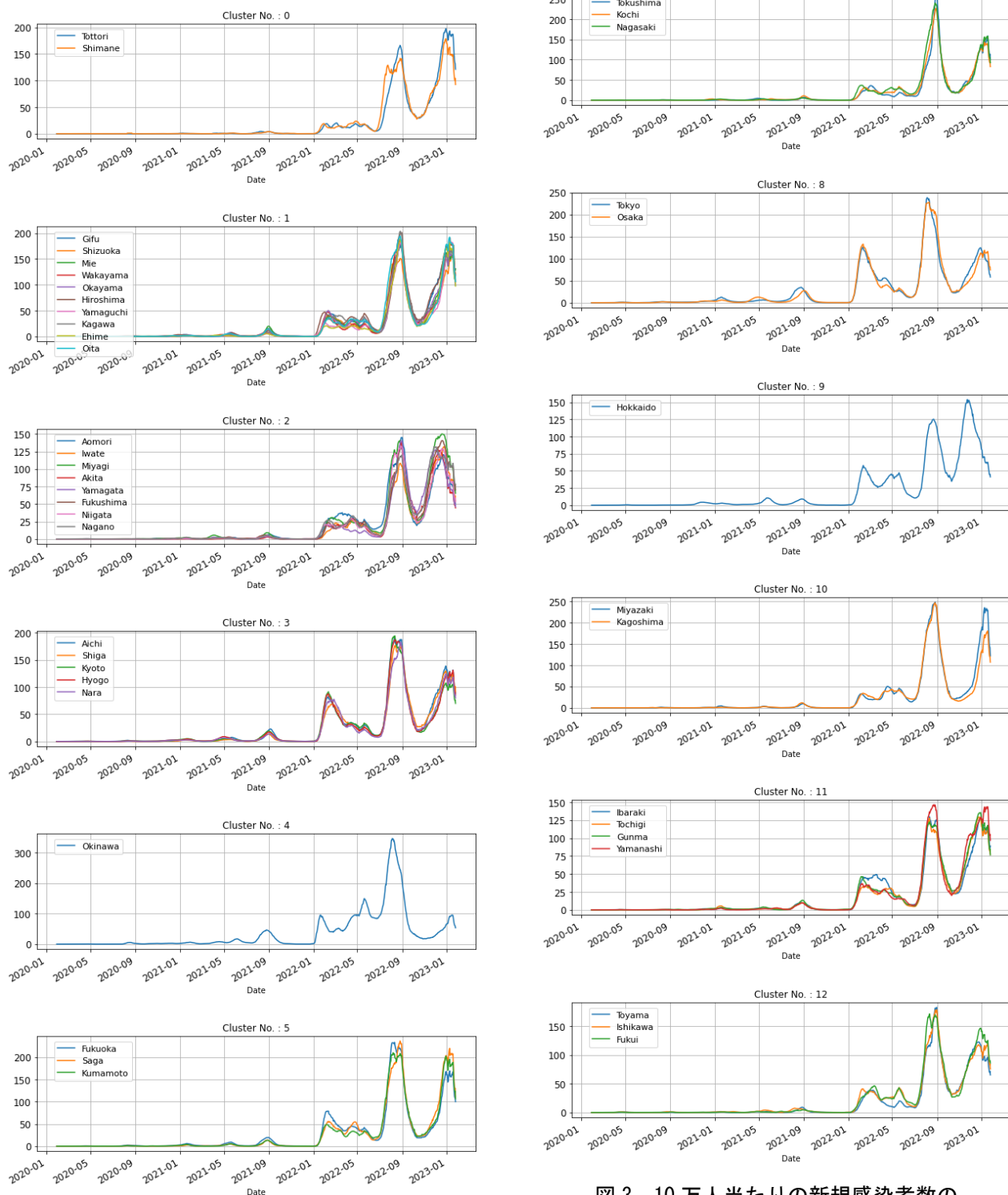


図3 10万人当たりの新規感染者数の14日平均の Time Series k-means 法によるクラスタリング (2020/1/16~2023/1/26)

前回報告と同様、これらのクラスター分類の裏付けとなる要因は直接的には述べられないが、距離的な近さ、生活圈や交通量、大都市、その周辺という地域性が関係しているものと推測される。

3. アジア諸国とヨーロッパ諸国についての分析

ここでは、2章と同様にアジア諸国とヨーロッパ諸国について分析する。データは世界保健機関（WHO）のサイト⁽²⁾から取得した2020/1/16～2023/1/26のものである。図4と図6は、それぞれアジア諸国間、及びヨーロッパ諸国間の相関係数のヒートマップとデンドログラムである。また図5と図7は、それぞれアジア諸国とヨーロッパ諸国の標準化された新規感染者数7日平均のTime Series k-Means法⁽³⁾によるクラスタリングの結果

である。クラスター数は、エルボー法などにより両方とも10とした。図4と図5からアジアでは、日本、中国の各単独クラスター、インドを中心とした近隣国クラスター（インド/ネパール（No. 6）、パキスタン/バングラデッシュ（No. 3））、遠距離だが半島国家同士のクラスター（韓国/ベトナム/シンガポール（No. 7））、マレー半島の隣接国クラスター（マレーシア/タイ（No. 0））、インド洋の沿岸国を含むクラスター（インドネシア/ミャンマー（No. 5）スリランカ/カンボジア/アフガニスタン（No. 8））、などが形成されている。今回追加した2021年7月以降のデータ内の大きな波は、それより前の波の数倍の規模であるが、同様の近隣国や地理的な条件が類似している国々から成るクラスターが構成されている。

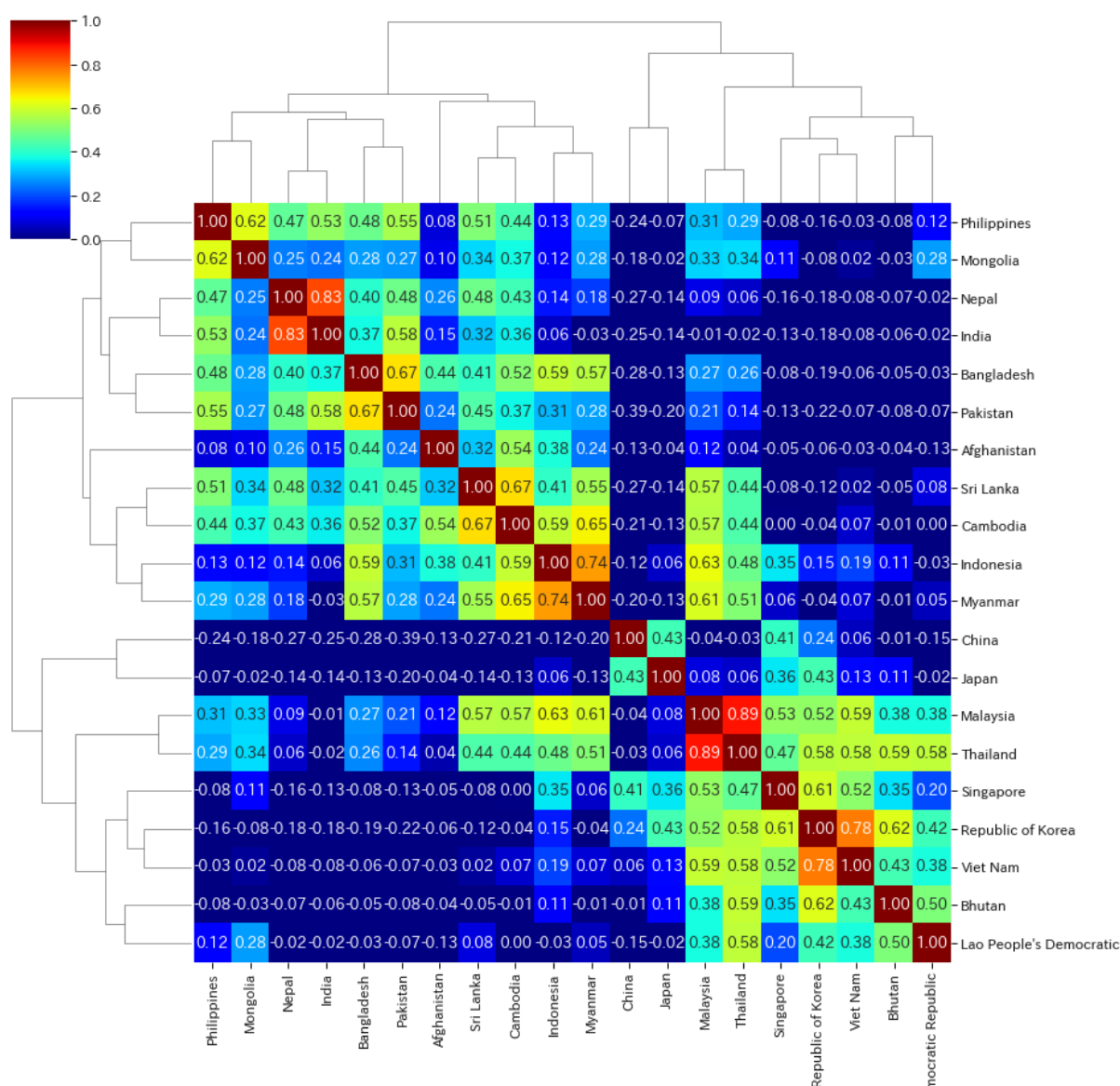


図4 アジア諸国間の相関係数のヒートマップとデンドログラム（2020/1/16～2023/1/26）

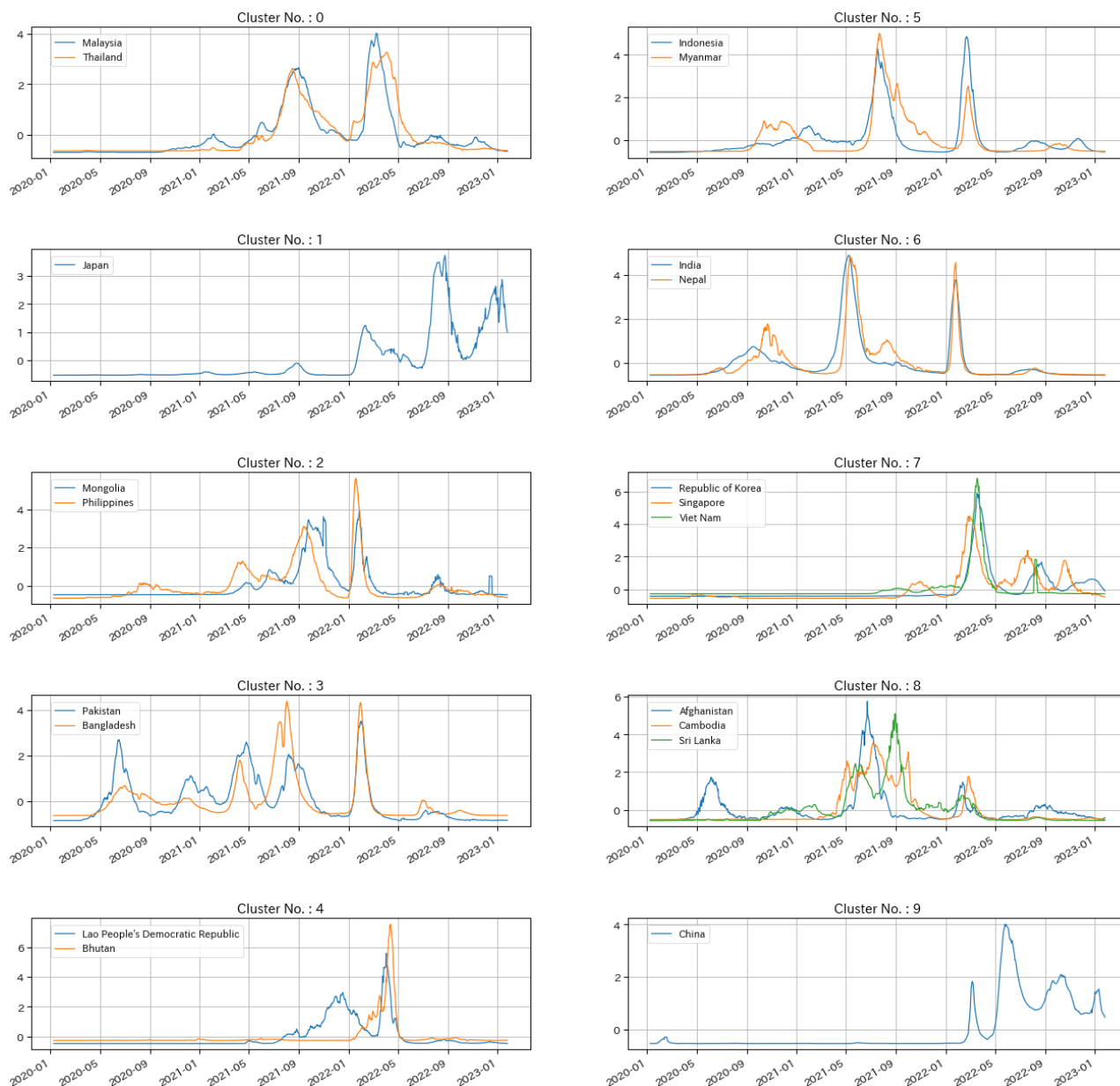


図5 アジア諸国の標準化新規感染者数7日平均のTime Series k-Means 法によるクラスタリング

図6と図7からヨーロッパでは、大西洋から地中海にかけての沿岸諸国のクラスター（フランス/イタリア/ルクセンブルグ/モナコ（No. 2）、アイルランド/スペイン/イギリス（No. 3）、ベルギー/ポルトガル/スウェーデン（No. 5））、北海の沿岸諸国と隣接国のクラスター（デンマーク/アイスランド/オランダ/ノルウェー/スイス（No. 1）、オーストリア/フィンランド/ドイツ（No. 4））、旧東欧諸国周辺のクラスター（ロシアとスロバキア（No. 7）、ハンガリーと北マケドニアとポーランド（No. 6）、ブルガリアとルーマニアとセルビア（No. 0））、アルバニア（No. 8）とギリシャ（No. 9）は、それぞれ単独のクラスターとなっている。

今回は、前回の報告で使用したデータに2021年7月以降のデータを追加して分析している。ヨーロッパにおける2022年1月頃の大きな波は追加データに含まれているものだが、それよりも前の波に比べ数倍の大きさになっている。本報告でも前回の報告と同様に、旧東欧諸国周辺のクラスターが形成されている。また、それ以外の国のクラスター構成は若干異なるものの、今回の報告でも全体として、この旧東欧諸国周辺のクラスター、大西洋から地中海にかけての沿岸諸国のクラスター、北海の沿岸諸国と隣接国のクラスター、という大きく分けて3つのクラスターが形成されていると推察される。

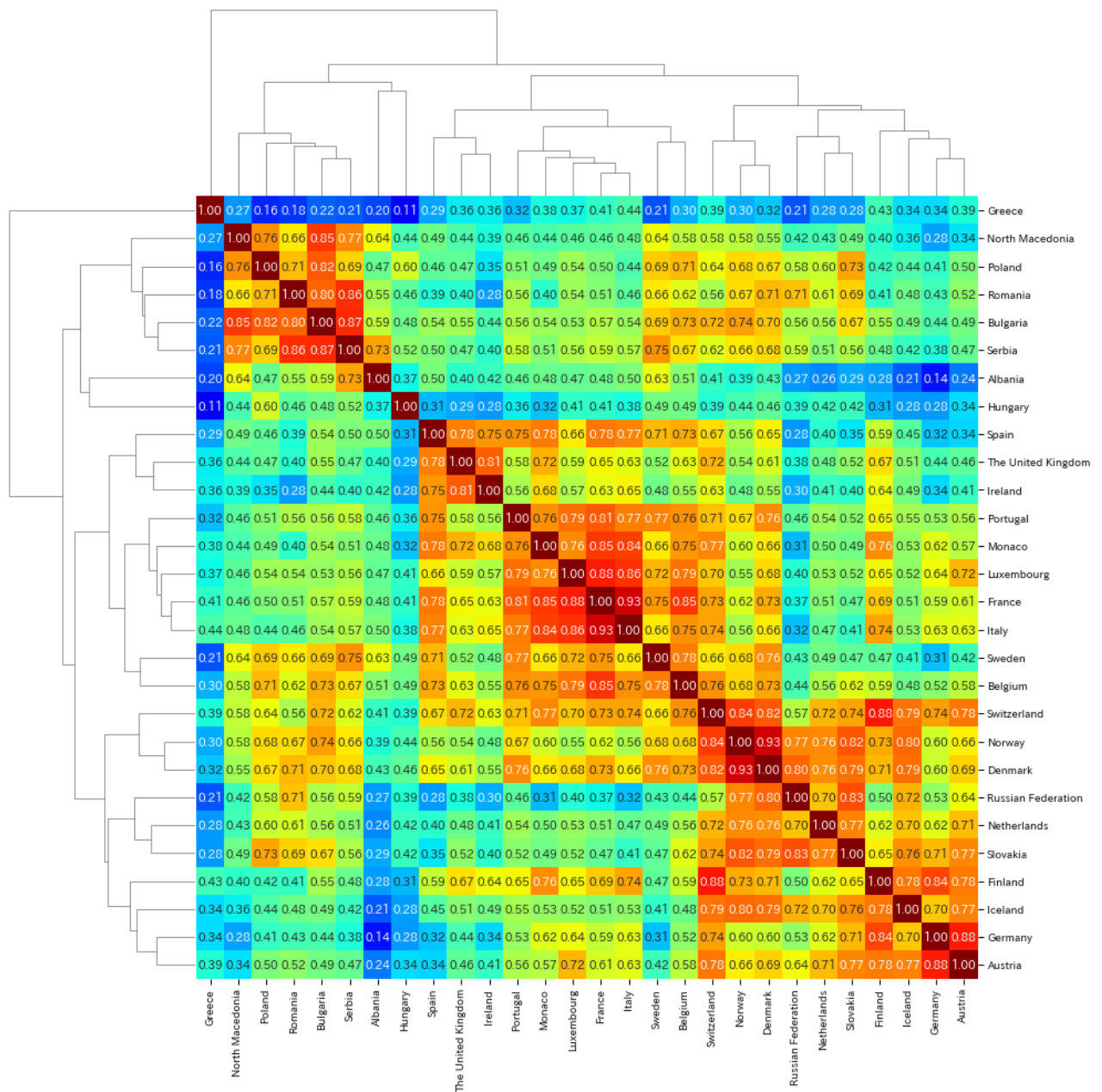


図 6 ヨーロッパ諸国間の相関係数のヒートマップとデンドログラム (2020/1/16~2021/6/28)

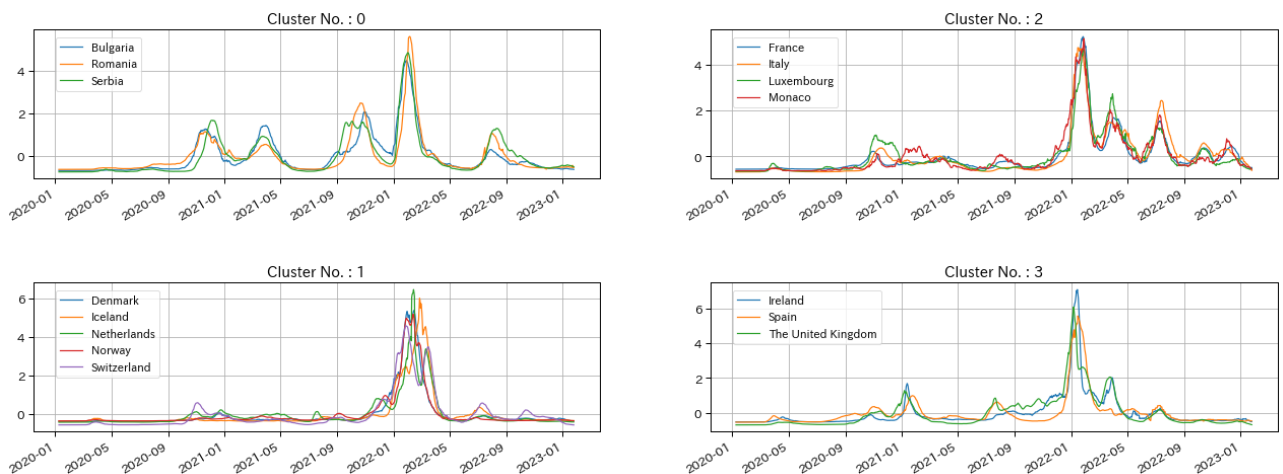




図7 ヨーロッパの標準化された新規感染者数7日平均のTime Series k-Means 法によるクラスタリング

4. まとめ

COVID-19 の新規感染者数のオープンデータを時系列データと捉えて、2022 年の新たなデータを追加して相関分析と機械学習の教師なし学習によるクラスタリングを試みた。追加されたデータでは、大きな波の規模が数倍になっているケースも多かったが、過去の報告と同様にデータサイエンスの視点からの分析におけるいくつかの知見は得られたと考えられる。今回の含む3回のクラスター分析により、隣接する地域や地理的に類似している地域、人で行き来や交通量が多い地域から構成されるクラスターが構成される傾向は確認できてきていると考えられる。

今後も、引き続き新たなデータを加えた分析や、他のデータサイエンスや機械学習の分析手法との比較、それらを組み合わせた分析や応用が必要であると考えられる。また、同様の感染拡大と縮小をするような感染症（例えば季節性のインフルエンザなど）の新規感染者数のデータを使った同様のクラスタリング分析を実施し、その結果と比較して類似したクラスターが形成されるかどうかなどを検証することも興味あるテーマであると考えられる。また、COVID-19 の感染が収束するまでに蓄積されるであろう多くのデータを即時に分析して公開するシステムの構築なども重要な課題であると考えられる。

参考文献

- (1) 厚生労働省：データからわかる－新型コロナウイルス感染症情報－，2023 年 1 月 26 日，<https://covid19.mhlw.go.jp/>。
- (2) World Health Organization: WHO Coronavirus (COVID-19) Dashboard, 2023 年 1 月 26 日，<https://covid19.who.int/info/>。
- (3) Xiaohui Huang, Yunming Ye, Liyan Xiong, Raymond Y.K. Lau, Nan Jiang, Shaokai Wang: Time series k-means: A new k-means type smooth subspace clustering for time series data, Information Sciences, Volumes 367-368, 1 November, pp. 1-13, 2016.
- (4) 機械学習を用いた COVID-19 新規感染者数の分析, 小松 隆行, 北海道科学大学研究紀要, (49), 49-61 (2021-09-30)。
- (5) 教師なし学習を用いた COVID-19 新規感染者数のクラスター分析, 小松 隆行, 北海道科学大学研究紀要, (50), 61-68 (2022-09-30)。